Uncovering Argumentative Flow: A Question-Focus Discourse Structuring Framework

Anonymous ACL submission

Abstract

Understanding the underlying argumentative flow in analytic argumentative writing is essential for discourse comprehension, especially in complex argumentative discourse such as think-tank commentary. However, existing structure modeling approaches often rely on surface-level topic segmentation, failing to capture the author's rhetorical intent and reasoning process. To address this limitation, we propose a Question-Focus discourse structuring framework that explicitly models the underlying argumentative flow by anchoring each argumentative unit to a guiding question (reflecting the author's intent) and a set of attentional foci (highlighting analytical pathways). To assess its effectiveness, we introduce an argument reconstruction task in which the modeled discourse structure guides both evidence retrieval and argument generation. We construct a high-quality dataset comprising 600 authoritative Chinese think-tank articles for experimental analysis. To quantitatively evaluate performance, we propose two novel metrics: (1) Claim Coverage, measuring the proportion of original claims preserved or similarly expressed in reconstructions, and (2) Evidence Coverage, assessing the completeness of retrieved supporting evidences. Experimental results show that our framework uncovers the author's argumentative logic more effectively and offers better structural guidance for reconstruction, yielding up to a 10% gain in claim coverage and outperforming strong baselines across both curated and LLM-based metrics.

1 Introduction

Analytical argumentative writing is a structured form of discourse, designed to dissect intricate issues, evaluate multiple perspectives, and articulate a well-founded position through systematic reasoning. The primary purpose is not merely to state opinions but to demonstrate the validity of a claim using well-supported evidence and logical connections. Central to this process is the concept of *argumentative flow*, which refers to the seamless progression of these components, ensuring that each section logically connects to the next. A well-executed argumentative flow enhances readability and persuasiveness, guiding the audience through the reasoning process without confusion. Whether in essays, debates, or research papers, mastering this flow is essential for constructing convincing and intellectually rigorous arguments.

Modeling such logic flow through discourse structure analysis has long been a foundational task in natural language processing (NLP)(Dijk and Kintsch, 1983), yet remains challenging due to the implicit and multi-layered nature of argumentative flow. Accurately uncovering this structure is essential for a range of downstream tasks, including document understanding (Chivers et al., 2022), information extraction (Aumiller et al., 2021; Xu et al., 2024a), question answering (Xu et al., 2024b), automatic writing (Liang et al., 2024; Gao et al., 2023; Shen et al., 2023), and controlled text generation (Fan et al., 2018; Rashkin et al., 2020; Fang et al., 2021; Li et al., 2023). However, most prior works (Koshorek et al., 2018; Arnold et al., 2019) rely on surface-level topic segmentation and hierarchical keyword outlines to represent discourse structure. While these coarse outlines provide a general overview, they often fail to capture the underlying argumentative flow—namely, *why* a section is written and *how* the author develops the argument (Asher, 2004). This gap is particularly critical in argumentative discourse modeling, as surface-level outlines cannot faithfully reconstruct the author's reasoning flow and rhetorical intent.

To fill this gap, we revisit the classical structure of argumentative discourse: each argumentative unit typically centers around a claim supported by one or more evidence. Crucially, what makes the reasoning persuasive is not the evidence

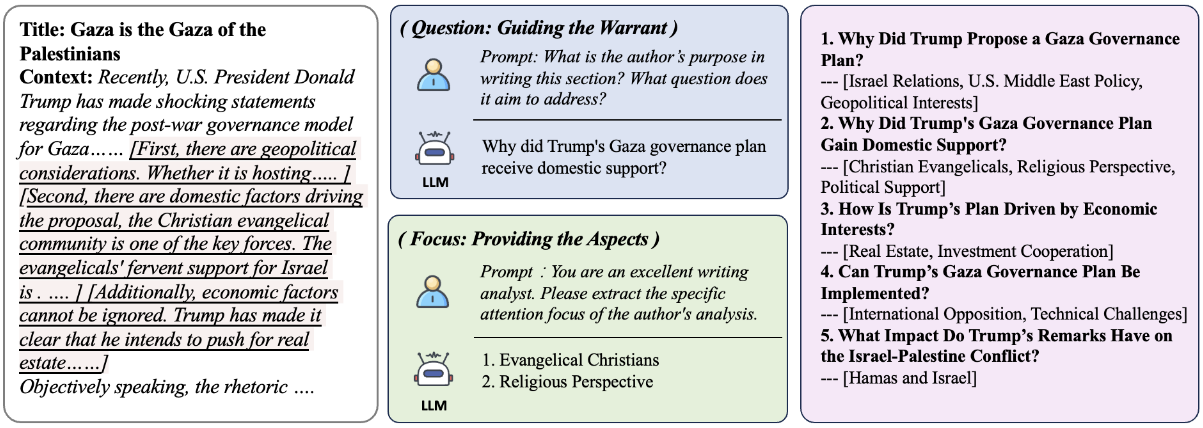

Figure 1: An example illustrating the proposed Question-Focus Discourse Structuring framework. The left panel shows the original article with its main argumentative section highlighted. The central panel presents the components of Question-Focus discourse structure: question (guiding the warrant) and attentional focus (providing the aspects). The right panel displays overall discourse frame, which structurally represents the article's argumentative flow through a sequence of question–focus pairs.

alone, but the underlying warrant—an implicit rationale that justifies why the premise supports the claim (Habernal et al., 2017). The warrant serves as a hidden bridge, encoding the author's reasoning process and shaping the reader's understanding of the argument. Additionally, we draw insight from the intentional structure theory proposed by Grosz and Sidner (Grosz and Sidner, 1986), which views discourse as a goal-driven process composed of linguistic sequencing, communicative intent, and attentional focus. This perspective highlights the importance of modeling why it is said and how it is developed. Given these considerations, modeling argumentative flow necessitates an explicit guiding component that not only directs the generation of appropriate warrants by aligning them with the author's communicative intent, but also determines the analytical focus—the specific aspects or dimensions through which the argument should be developed.

Building on these insights, we propose a *Question-Focus* discourse structuring framework to uncover the underlying argumentative flow for analytic argumentative writing. The main component of framework consists of a guiding question and a set of attentional foci for each argumentative unit. As shown in Figure 1, with the strong capabilities of LLM (Zhao et al., 2023; Chang et al., 2024), for an argumentative unit that analyzes U.S. domestic support for Trump's Gaza policy, the generated guiding question—" Why did Trump's Gaza governance plan receive domestic support? "—not only clarifies the author's argumentative intent but also implicitly surfaces the warrant: religious identity shapes political alignment. Moreover, the attentional foci, such as " Evangelical Christians" and "Religious Perspective", further highlight the reasoning emphasis. Together, these elements form a structured discourse frame that models the author's reasoning trajectory.

To assess the effectiveness of our discourse structuring framework, we introduce an argument reconstruction task that simulates human-like writing of persuasive argumentative articles. This task is structured in two phases: first, retrieving contextually relevant evidence guided by the hierarchical discourse structure, and second, synthesizing argument units that align with the pre-defined organizational schema. To facilitate this evaluation, we construct a dataset comprising 600 high-quality argumentative articles sourced from authoritative Chinese think tanks for experimental validation. To quantitatively evaluate performance, we introduce two novel metrics: *claim coverage* and *evidence coverage*, which measure the degree to which reconstructed arguments preserve the key elements of the original texts. These metrics not only assess fidelity to the source material but also illuminate how effectively our Question-Focus discourse structure directs the argument regeneration process. Our experimental results reveal that the proposed framework demonstrates superior capability in capturing authentic argumentative flow, achieving significant improvements over competitive baseline methods across both curated metrics and LLM-based assessments.

2 Related Work

2.1 Discourse Structure Modeling

Document structure modeling seeks to capture the internal organization of long-form texts. A common approach is to segment the document into coherent units and generate section headings to reveal its content structure—a process known as outline generation(Zhang et al., 2019; Inan et al., 2022; Barrow et al., 2020). Such topic-based hierarchical representations have been widely applied in expository genres such as Wikipedia articles and scientific writing(Fan and Gardent, 2022; Shao et al., 2024), as well as in noisier domains like meeting transcripts or podcast recordings, where outlines serve to impose post-hoc structure onto otherwise unstructured content(Retkowski and Waibel, 2024; Ghazimatin et al., 2024).In narrative or story-centric documents, document structure is often modeled through event sequences or temporal plots, rather than thematic section headers, reflecting the underlying causal or chronological structure of the text(Fang et al., 2021; Li et al., 2023). Beyond hierarchical topic modeling, some work has explored using summary-level representations—such as paragraph-level abstractive summaries—as an alternative structure to guide document understanding or generation(Sun et al., 2020). Despite recent progress, most methods rely on uniform, topic-based outlines built from surface cues, overlooking genre-specific discourse structures. Large language models (LLMs), with their strong semantic understanding and generative capabilities, offer new potential for modeling document structures beyond simple topic segmentation, enabling more nuanced and genre-aware representations(Zhao et al., 2023).

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances language models (LMs) by retrieving external information at inference time to improve factuality and informativeness(Ram et al., 2023; Izacard et al., 2023). Existing work mainly explores two directions: one uses retrieved texts as in-context examples to guide generation(Li et al., 2023; Liu et al., 2021; Agrawal et al., 2022; Poesia et al., 2022; Khattab et al., 2022; Shi et al., 2022), while the other incorporates retrieved evidence directly into the input to ground the output and reduce hallucinations(Lewis et al., 2020; Semnani et al., 2023). Despite growing interest in RAG, its application to long-form article generation remains underexplored. RAG has been widely applied to tasks like question answering, dialogue, and citation-based generation(Menick et al., 2022; Gao et al., 2023; Bohnet et al., 2022). It also supports flexible retrieval sources, ranging from domain-specific databases (e.g., medicine, finance) to open-domain web content(Zhou et al., 2022; Nakano et al., 2021) and code documentation(Zakka et al., 2024).

3 Methods

We propose a Question-Focus Discourse Structuring approach with LLMs to capture the underlying logical flow of argumentative discourse (§3.1). To validate its effectiveness, we introduce an argument reconstruction task that simulates expert writing through evidence retrieval and structured argument generation over full-length argumentative articles (§3.2.1–§3.2.2). Figure 2 provides an overview of our framework.

3.1 Question-Focus Discourse Structuring

A well-structured writing plan is widely acknowledged to be critical for producing coherent and high-quality texts(Sun et al., 2020; Yang et al., 2022b,a), especially in argumentative discourse, where clarity of reasoning between claims and premises is crucial. Inspired by the role of warrants—the implicit justifications linking premises to claims(Habernal et al., 2017)—and the intentional structure theory(Grosz and Sidner, 1986), we propose a cognitively grounded Question-Focus Discourse Structuring approach. Each argumentative unit is anchored by a guiding question that captures the author's rhetorical intent and implicitly guides the underlying warrant. We also extract attentional foci, the key analytical aspects emphasized in the reasoning. Together, these elements form a structured representation of the author's argumentative flow, enabling interpretable and structure-aware generation.

We design a three-stage, LLM-assisted pipeline to model the discourse structure of full-length argumentative articles. First, given an input document D, we prompt the LLM to segment it into a sequence of fine-grained *argumentative units* $\{AU_1, AU_2, \ldots, AU_n\}$, each representing a self-contained block of reasoning that contributes to the overall argumentative progression (Figure 2 1). Concurrently, the LLM extracts contextual metadata, including the topic T, core problem P, and background information B, which provide global

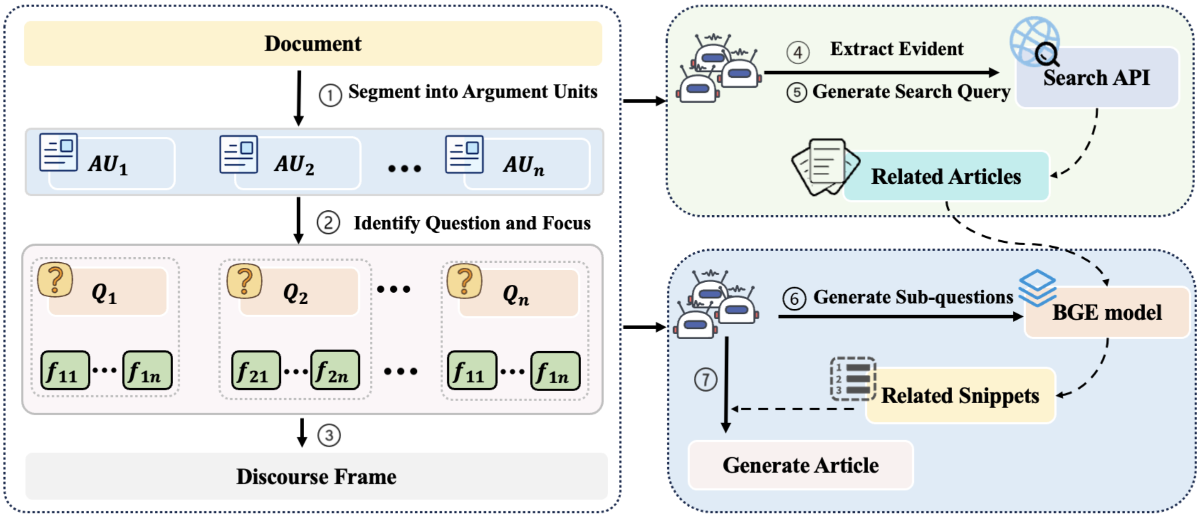

Figure 2: Overview of our framework. Steps(1-3) construct a question-focus discourse structure by identifying argumentative units, guiding questions, and attentional foci. Steps(4-5) retrieve external evidence via LLM-based strategies, extracting factual claims and generating queries guided by the discourse structure. Steps(6-7) perform argument reconstruction by decomposing each guiding question into sub-questions, retrieving relevant evidence snippets, and generating grounded content.

guidance for subsequent modeling. Next, for each argumentative unit AU_i, the LLM is prompted to infer a *guiding question* Q_i that captures the author's rhetorical intent and serves to guide the underlying reasoning strategy (Figure 2 ②). This guiding question implicitly reflects the warrant, which frames how the premise supports the claim. In parallel, we extract a set of *attentional foci* $f_i = \{f_{i1}, f_{i2}, \dots, f_{im}\}$, which represent the key analytical aspects emphasized in answering Q_i. Finally, we compose all units into a structured *discourse frame* $F = (AU_1, Q_1, f_1), ..., (AU_n, Q_n, f_n)$ (Figure 2 ③), which captures the argumentative flow, communicative intent, and focal perspectives of the article. This frame provides a cognitively grounded foundation for structure-aware argument reconstruction.

3.2 Argument Reconstruction

Given the extracted question-focus discourse structure, we simulate an expert writing process, which comprises evidence collection and argument generation, with the assistance of LLMs.

3.2.1 Evidence Collection

Argumentative articles typically lack explicit citations, making it challenging to trace their underlying sources. To address this, we adopt a dual-faceted retrieval strategy. First, for each segment AU_i in the article, we prompt the LLM to extract factual assertions C_i^{fact} from the source text (Figure 2 ④), which serve as implicit evidence cues for locating original or semantically related documents. Second, leveraging the structured discourse frame $F = \{(AU_i, Q_i, f_i)\}$ and contextual metadata (T, P, B), the LLM generates guided search queries C_i^{struct} based on each unit's guiding question Q_i and attentional focus f_i (Figure 2 ⑤). The combined set of queries $C_i = C_i^{\text{fact}} \cup C_i^{\text{struct}}$ is submitted to the Tavily Search API[1] to retrieve relevant external articles. This evidence retrieval process helps ground the subsequent generation in relevant facts and increases the likelihood of recovering the author's original argumentative stance."

3.2.2 Article Generation

Building on the retrieved references R and the structured discourse frame $F = \{(AU_i, Q_i, f_i)\}$, we reconstruct the article in a unit-wise manner. For each argumentative unit AU_i, the LLM is prompted to generate a set of sub-questions $\{q_{i1}, q_{i2}, \dots\}$, derived from its guiding question Q_i, attentional focus f_i, and the metadata (T, P, B) (Figure 2⑥). Since it is typically infeasible to include the entire reference set R within the LLM's context window, we use these sub-questions to retrieve semantically relevant evidence snippets from R, based on BGE-based Sentence-BERT embeddings. The LLM then generates the content of AU_i, grounded in the retrieved evidence (Fig-

[1] https://tavily.com/

ure 2(7)). As all units are reconstructed independently, we finally prompt the LLM to generate the introduction and conclusion using the global metadata (T, P, B), ensuring overall coherence of the reconstructed article.

4 Experiments

4.1 Dataset

Despite recent progress in LLM-assisted expository and narrative writing (Shao et al., 2024; Lee et al., 2024; Yang et al., 2022b), the domain of argumentative discourse, including think tank commentaries, remains largely underexplored. The lack of high-quality datasets in this area limits the development and evaluation of structure-aware generation methods for real-world argumentative writing. To fill this gap, we curate a dataset of 600 high-quality argumentative articles, carefully selected from authoritative Chinese think tanks, including the China Institute of International Studies[2] . These articles span a broad range of global issues and are authored by domain experts. Each article presents a well-defined argumentative structure, including explicit claims, supporting evidence, and in-depth reasoning informed by expert analysis. Given that our target texts (e.g., think tank commentary and policy analysis) are typically unstructured plain text without section headings or references, our dataset consists entirely of such free-form discourse. This provides a valuable foundation for discourse structure modeling and structure-aware generation tasks such as argument reconstruction.

4.2 Metrics

To assess whether our question-focus discourse structure effectively guides LLMs in reconstructing argumentative texts, we adopt a combination of custom-designed and standard evaluation metrics that jointly evaluate semantic alignment, factual consistency, and overall content quality.

Argumentative writing is centered on conveying the author's viewpoints through structured reasoning and evidence (Wenzel et al., 1992; Qin and Liu, 2021). To evaluate how well the reconstructed argument preserves these original intentions, we introduce two claim-level metrics: **Claim Coverage Rate (CCR)** and **Claim Entity Recall (CER)**. CCR quantifies the semantic similarity between core claims extracted from the human-written article (considered as ground truth) and those extracted from the reconstructed text, using Sentence-BERT embeddings (Chen et al., 2024) (details in Appendix A.1). CER measures the percentage of named entities in the ground-truth claims that appear in the reconstructed set, using the LAC named entity recognition (NER) toolkit (Jiao et al., 2018). To further assess factual consistency, we introduce **Evidence Coverage Rate (ECR)** : this metric calculates how well the reconstructed argumentative units recover the factual content found in the original article (details in Appendix A.1).

For overall article quality, we report ROUGE scores (Lin, 2004) and entity recall over the full article, providing auxiliary indicators of textual overlap and factual completeness. Furthermore, we prompt two advanced LLMs, GPT-4o (Hurst et al., 2024) and DeepSeek-R1 (Guo et al., 2025), to evaluate each reconstructed article relative to its original across five key dimensions: *Relevance*, *Structure*, *Coverage*, *Accuracy*, and *Coherence*, using a 5-point rubric (Kim et al., 2023) (see Appendix A.2).

4.3 Baselines

Modeling the discourse structure of argumentative texts with LLMs remains largely underexplored. A closely related task is *outline generation* (Zhang et al., 2019; Barrow et al., 2020; Jiang et al., 2023; Inan et al., 2022), which also aims to capture the underlying structure of a document. In many LLM-based automatic writing systems(Shao et al., 2024; Lee et al., 2024), a content plan is first constructed before full-text generation. These plans typically take the form of hierarchical outlines composed of short section and subsection titles, which serve as coarse-grained signals to guide subsequent content generation. We refer to such structures as **rough outlines**. Other studies have proposed more fine-grained content planning strategies by providing sentence-level(Li et al., 2023) or summary-level outlines(Sun et al., 2020).

However, prior works use traditional setups and do not use LLMs. As such, they are difficult to compare directly with our framework. Instead, to establish fair and meaningful comparisons, we adapt representative ideas from existing work and design the following three LLM-based baselines:

Rough-direct This baseline represents the dominant paradigm in current LLM-based writing systems. The model first segments the article and generates a coarse hierarchical outline based on

[2] https://www.ciis.org.cn/

Model	Method	ROUGE - 1	ROUGE - 2	ROUGE - L	Entity Recall
GPT-3.5	Rough-Direct	30.40	8.13	17.27	19.31
	Rough-RAG	33.12	10.14	17.88	24.98
	SOE	36.26	11.75	18.46	24.54
	Question-Focus	**44.96**	**17.64**	**21.71**	**50.34**
	w/o focus	35.07	11.00	18.10	26.55
GPT-4	Rough-Direct	29.86	8.01	17.17	19.26
	Rough-RAG	33.53	10.97	18.42	25.49
	SOE	37.17	12.19	18.44	26.11
	Question-Focus	**49.76**	**24.10**	**24.99**	**53.55**
	w/o focus	33.89	10.70	17.92	26.42
DeepSeek-V3	Rough-Direct	29.67	7.04	14.82	28.83
	Rough-RAG	31.60	8.35	14.71	30.78
	SOE	35.07	10.55	16.84	34.37
	Question-Focus	**47.39**	**20.95**	**23.19**	**57.22**
	w/o focus	34.59	10.98	17.93	30.81

Table 1: Comparison of different models on article reconstruction, evaluated against human-written articles using ROUGE-1, ROUGE-2, ROUGE-L, and Entity Recall (%). Bold values indicate the best performance.

high-level topics (typically expressed as keywords or short phrases), and then directly generates the reconstructed text conditioned on this outline. This structure-first pipeline has been widely adopted in expository writing , such as Wikipedia generation (Shao et al., 2024). We include this baseline to evaluate how well such a commonly used yet coarse structural representation performs in reconstructing argumentative articles.

Rough-RAG This baseline extends Rough-Direct by incorporating retrieval-augmented generation (RAG) in the reconstruction phase. As RAG techniques have become increasingly popular for enhancing the factual accuracy of LLM outputs(Lewis et al., 2020), the LLM is guided by the outline while retrieving and incorporating relevant external evidence from online sources.

SOE This baseline adopts the Summarize-Outline-Elaborate (SOE) method proposed by Sun et al. (2020), which models fine-grained argumentative logic through summary-based planning. The process first segments the input article into coherent discourse units. For each unit, the model generates a concise summary that captures its core idea. These summaries are then organized into a structured outline representing the article's overall argumentative flow. The LLM then reconstructs the full article by elaborating each unit based on its summary, aiming to preserve the original intent and logical structure.

4.4 Implementation Details

We implement our pipeline in two main stages: question-focus discourse structuring and argument reconstruction, using zero-shot prompting within the DSPy framework (Khattab et al., 2023). Appendix B includes the pseudo code and corresponding prompts. For the discourse structuring stage, including document segmentation and metadata extraction, we use the open-source `Qwen2.5-7B-Instruct` model, deployed on an NVIDIA A800 GPU, with a default top_p setting of 0.8. For guiding question generation, attentional focus extraction, and argument reconstruction, we experiment with `gpt-3.5-turbo`, `gpt-4`, and `deepseek-V3`. In the argument reconstruction stage, we retrieve external evidence using the Tavily Search API[3], excluding the original article from the retrieval pool to avoid data leakage. The pipeline remains compatible with other search engines. For all LLM-based generation steps (except Qwen), we set the temperature to 1.0 and the top_p value to 0.9.

5 Results and Analysis

5.1 Analysis of Claim-Evidence Coverage

We evaluate the effectiveness of our proposed framework in argument reconstruction using three targeted metrics: Claim Coverage Rate (CCR), claim Entity Recall (CER), and Evidence Coverage Rate (ECR) (see §4.2). These metrics collectively assess how well the reconstructed article preserves

[3] https://www.tavily.com

Model	Method	**Relevant**	**Structure**	**Coverage**	**Accuracy**	**Coherence**	**Overall**
GPT-4o	Rough-Direct	3.18	2.55	2.29	3.24	3.20	3.15
	Rough-RAG	3.72	3.22	2.84	3.76	3.59	3.65
	SOE	3.95	3.47	3.01	4.07	3.73	3.86
	Question-Focus	**4.43†**	**3.53**	**3.7†**	**4.55†**	**4.33†**	**4.32**
	w/o focus	3.72	3.23	2.99	3.79	3.62	3.69
DeepSeek-R1	Rough-Direct	3.05	2.74	2.56	3.09	3.39	3.04
	Rough-RAG	3.56	3.36	3.24	3.41	3.69	3.53
	SOE	3.83	3.32	3.39	3.96	3.83	3.79
	Question-Focus	**3.95**	**3.71**	**3.52**	**4.51**	**4.34**	**4.2**
	w/o focus	3.50	3.32	3.32	3.66	3.7	3.56

Table 2: LLM-based evaluation results across five dimensions: Relevance, Structure, Coverage, Accuracy, and Coherence. Bold indicates the highest score;, and † denotes significant improvements over all baselines.The rubric grading uses a 1-5 scale.

the author's intent, argumentative content, and factual grounding. As shown in Table 3 and Table 4, our method consistently outperforms all baselines across GPT-3.5, GPT-4, and DeepSeek-V3 backbones. Notably, we achieve the highest scores on all models—e.g., on GPT-4, CCR/CER reach 86.18 / 79.13, and ECR reaches 86.58. Compared to the strongest baseline *SOE*, our approach yields gains of up to +11.07 in CCR, +5.75 in CER, and +11.32 in ECR, demonstrating its superior ability to recover both the author's viewpoints and supporting evidence.

Among the baselines, *Rough Direct*, which uses only coarse hierarchical outlines, shows moderate CCR (60–70%), indicating that LLMs can leverage their rich parametric knowledge in combination with surface-level structure to partially recover central claims. *Rough-RAG* improves upon this by incorporating retrieved external knowledge, validating the importance of evidence grounding. Notably, *SOE*, which builds from sentence-level summaries, captures more focused argumentative content and yields stronger performance across metrics. Nevertheless, our method still surpasses *SOE*, showing that explicitly modeling the discourse structure with question–focus pairs not only provides finer-grained rhetorical control, but also enhances interpretability and fidelity by aligning generation with the original argumentative flow.

5.2 Analysis of Reconstruction Quality

We further assess the quality of reconstructed articles by directly comparing them to their human-written counterparts. As shown in Table 1, our method consistently outperforms baselines on ROUGE metrics and Entity Recall. Compared to the strongest baseline *SOE*, our method improves ROUGE-1 by up to +12.59, ROUGE-2 by +11.91, ROUGE-L by +6.55, and Entity Recall by +27.44 , indicating a higher degree of content fidelity and textual alignment. *Rough-RAG* shows improvements over *Rough-Direct* by integrating external evidence, while *SOE* benefits from summary-level structuring. However, our approach, which combines question-focus discourse structuring with structure-guided generation, achieves markedly superior results, underscoring the effectiveness of explicitly modeling argumentative flow to guide faithful reconstruction.

5.3 LLM-Based Evaluation

Table 2 presents LLM-based evaluation results across five key dimensions—*Relevance*, *Structure*, *Coverage*, *Accuracy*, and *Coherence*—along with an overall quality score. Our method achieves the highest ratings across all dimensions, especially excelling in relevance, information coverage, accuracy and coherence, demonstrating its effectiveness in preserving the original article's argumentative logic and factual content. The overall quality score further confirms the superiority of our approach in generating coherent and faithful reconstructions. Additionally, evaluations by two distinct LLMs (GPT-4o and DeepSeek-R1) show minimal variance (within 0.5 points), indicating strong robustness across evaluation settings.

Taken together, **our question-focus discourse structuring and guided reconstruction approach yields significant gains in content fidelity and alignment with the author's reasoning. By explicitly modeling the argumentative flow, it enables more faithful and interpretable reconstruction, consistently outperforming all baseline methods.**

		CCR	CER
GPT-3.5	Rough Direct	68.69	65.01
	Rough-RAG	71.60	68.46
	SOE	74.88	71.83
	question-focus	**82.65†**	**76.83†**
	w/o focus	80.06	75.63
GPT-4	Rough Direct	71.83	67.19
	Rough-RAG	76.44	73.43
	SOE	75.11	73.38
	question-focus	**86.18†**	**79.13†**
	w/o focus	81.93	76.70
DeepSeek-V3	Rough Direct	62.26	68.63
	Rough-RAG	65.16	70.86
	SOE	66.78	72.78
	question-focus	**80.13†**	**77.54†**
	w/o focus	74.14	75.66

Table 3: Results of claim-level quality evaluation (%). Claim Coverage Rate (CCR) and Claim Entity Recall (CER) are computed based on LLM-extracted core claims from the original and reconstructed texts, assessing how well the reconstruction preserves the author's intended arguments. Bold values denote the best performance; † indicates significant improvement over all baselines.

	RD	RR	SOE	Ours	w/o f
ECR	58.41	70.26	75.26	**86.58**	79.86

Table 4: Results of Average factual quality (ECR, %) across different methods. Evidence Coverage Rate (ECR) measures how well the reconstructed article recovers factual content from the original. Bold values denote the best performance; RD(Rough-Direct), RR(Rough-RAG), w/o f (our model without attentional focus).

5.4 Ablation Studies

As described in Section §3.1, our framework models argumentative dicourse using a structured representation in which each argumentative unit is anchored by a guiding question and its corresponding attentional foci. To assess the contribution of the *focus* component, we conduct an ablation study by removing the foci and retaining only the guiding questions (*w/o focus*). In this setting, the reconstruction process is still directed by question-based intent modeling, but lacks explicit signals regarding the author's emphasis within each unit.

As shown in Tables 1, 2, 3, and 4, the full question–focus framework achieves the highest performance across all evaluation metrics, highlighting the critical role of attentional focus in discourse structuring and its downstream impact on argument reconstruction. We further examine the effectiveness of the guiding question alone. Results in Table 3 and Table 4 demonstrate that using only guiding questions (i.e., *w/o focus*) still outperforms the *SOE* baseline in CCR, CER, and ECR metrics. This suggests that guiding questions serve as effective anchors for inferring implicit warrants, enabling clearer modeling of argumentative flow and providing stronger guidance for faithfully reconstructing the author's reasoning.

6 Conclusion

We propose a question-focus discourse structuring framework that leverages LLMs to uncover the underlying logic flow of argumentative discourse. By modeling each discourse segment with guiding questions and attentional focus, our method provides an interpretable representation of the author's intent and reasoning trajectory.To evaluate its effectiveness, we introduce an argument reconstruction task and construct a high-quality think-tank article dataset, along with tailored evaluation metrics.Experiments show that our framework substantially improves the reconstruction quality, yielding better alignment with the original argumentative logic and content. These findings demonstrate the effectiveness of question-focus structuring in modeling complex argumentation. In future work, we plan to extend this framework to broader domains and explore its applications in interactive writing support and automated document planning.

Limitations

In this work, while our question-focus discourse structuring framework effectively guides argument reconstruction with superior performance across various automatic metrics. It is primarily validated on think-tank–style argumentative discourse with relatively clear segment-to-intent mappings. Nevertheless, our framework is inherently flexible and can be extended to handle more complex argumentative texts involving overlapping or evolving intents—such as by supporting multiple guiding questions and dynamic focus modeling within a single discourse unit. We leave this as a promising direction for future work.

Additionally, our reconstruction strategy uses retrieval-augmented generation to enhance factual grounding and reduce hallucination. However, sourcing evidence from the web inevitably introduces variability: online content may be time-sensitive, inconsistent, or factually unreliable , potentially affecting the accuracy and stance of the reconstructed argument. Moreover, different retrieved sources may present divergent analytical perspectives on the same guiding question. Although our pipeline incorporates a fact extraction step from the original article to guide retrieval and mitigate such risks, challenges in evidence verifiability remain. These verifiability issues go beyond typical hallucination concerns and point to broader challenges in ensuring source reliability for grounded text generation.

Ethics Statement

Our research focuses on argumentative articles such as think-tank commentaries, which serve as a key source of information for the public. All data used in our experiments are publicly available think-tank articles from authoritative sources. During the argument reconstruction process, online retrieval is conducted through publicly accessible APIs, and the retrieved content is used solely for research purposes.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A Gers, and Alexander Löser. 2019. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Nicholas Asher. 2004. Discourse topic. *Theoretical linguistics*, 30(2-3):163–201.

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 2–11.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas W Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Brian Chivers, Mason P Jiang, Wonhee Lee, Amy Ng, Natalya I Rapstine, and Alex Storer. 2022. Ants: a framework for retrieval of text segments in unstructured documents. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 38–47.

Teun Adrianus Van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press, New York, NY.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Angela Fan and Claire Gardent. 2022. Generating full length wikipedia biographies: The impact of gender bias on the retrieval-based generation of women biographies. *arXiv preprint arXiv:2204.05879*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Outline to story: Fine-grained controllable story generation from cascaded events. *arXiv preprint arXiv:2101.00822*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Azin Ghazimatin, Ekaterina Garmash, Gustavo Penha, Kristen Sheets, Martin Achenbach, Oguz Semerci, Remi Galvez, Marcus Tannenberg, Sahitya Mantravadi, Divya Narayanan, and 1 others. 2024. Podtile: Facilitating podcast episode browsing with auto-generated chapters. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4487–4495.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hakan Inan, Rashi Rungta, and Yashar Mehdad. 2022. Structured summarization: Unified text segmentation and segment labeling as a generation task. *arXiv preprint arXiv:2209.13759*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2023. Advancing topic segmentation and outline generation in chinese texts: The paragraph-level topic representation, corpus, and benchmark. *arXiv preprint arXiv:2305.14790*.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese lexical analysis with deep bi-gru-crf network. *arXiv preprint arXiv:1807.01882*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. *arXiv preprint*, arXiv:1803.09337.

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2024. Navigating the path of writing: outline-guided text generation with large language models. *arXiv preprint arXiv:2404.13919*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yunzhe Li, Qian Chen, Weixiang Yan, Wen Wang, Qinglin Zhang, and Hari Sundaram. 2023. Advancing precise outline-conditioned text generation with task duality and explicit outline control. *arXiv preprint arXiv:2305.14459*.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and 1 others. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Y Liu, D Iter, Y Xu, S Wang, R Xu, and C Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arxiv 2023. *arXiv preprint arXiv:2303.16634*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Gabriel Poesia, Oleksandr Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*.

Jingjing Qin and Yingliang Liu. 2021. The influence of reading texts on l2 reading-to-write argumentative writing. *Frontiers in Psychology*, 12:655601.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

Fabian Retkowski and Alexander Waibel. 2024. From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. *arXiv preprint arXiv:2402.17633*.

Sina J Semnani, Violet Z Yao, Heidi C Zhang, and Monica S Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *arXiv preprint arXiv:2305.14292*.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv:2304.02623*.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265.

Xiaofei Sun, Zijun Sun, Yuxian Meng, Jiwei Li, and Chun Fan. 2020. Summarize, outline, and elaborate: Long-text generation via hierarchical supervision from extractive summaries. *arXiv preprint arXiv:2010.07074*.

Joseph W Wenzel, William L Benoit, Dale Hample, and Pamela J Benoit. 1992. Perspectives on argument. *Readings in argumentation*, pages 121–143.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024a. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024b. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2022a. Doc: Improving long story coherence with detailed outline control. *arXiv preprint arXiv:2212.10077*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022b. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2019. Outline generation: Understanding the inherent content structure of documents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 745–754.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Shuyan Zhou, Uri Alon, Frank F Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2022. Docprompting: Generating code by retrieving the docs. *arXiv preprint arXiv: 2207.05987*.

A Automatic Evaluation Details

A.1 Claim and Evidence Coverage Rate

To assess whether the reconstructed article faithfully preserves the author's intended argumentative content, we define two semantic-level metrics: **Claim Coverage Rate (CCR)** and **Evidence Coverage Rate (ECR)**. Both metrics measure how well the reconstructed content semantically covers key claim or evidence units from the human-written article(treated as ground truth).

Let

$$O_{\text{ref}} = \{o_1^{\text{ref}}, o_2^{\text{ref}}, \ldots, o_m^{\text{ref}}\} \tag{1}$$

denote the set of core claims extracted from the human-written article, and

$$O_{\text{gen}} = \{o_1^{\text{gen}}, o_2^{\text{gen}}, \ldots, o_n^{\text{gen}}\} \tag{2}$$

the set extracted from the reconstructed article. All claims are obtained via LLM-guided prompts designed to elicit key propositions from each argumentative unit.

We compute the semantic similarity between each o_i^{ref} and all o_j^{gen} using cosine similarity over Sentence-BERT embeddings (we use the BGE model (Chen et al., 2024)). A reference claim is considered covered if its maximum similarity with any generated claim exceeds a threshold τ.

The CCR is calculated as:

$$\text{CCR} = \frac{1}{|O_{\text{ref}}|} \sum_{i=1}^{|O_{\text{ref}}|} \mathbb{I}\left[\max_j \text{sim}(o_i^{\text{ref}}, o_j^{\text{gen}}) > \tau\right] \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity function and $\mathbb{I}[\cdot]$ is the indicator function.

Evidence Coverage Rate (ECR) is computed analogously by replacing the claim sets with sets of factual evidence units extracted from the original and reconstructed articles:

- E_{ref} denotes evidence extracted from the human-written article.
- E_{gen} denotes evidence extracted from the reconstructed article.

The same computation method is applied, measuring the proportion of factual assertions from E_{ref} that are semantically matched by E_{gen}.

Both claim and evidence are extracted using LLM prompts. While claim prompts target subjective or evaluative viewpoints, evidence prompts are designed to identify verifiable factual assertions supporting those claims. See Appendix B for example prompts.

A.2 LLM evaluator

Recently, using powerful proprietary Large Language Models (LLMs) (e.g., GPT-4) as evaluators for long-form responses has become the de facto standard, due to their strong alignment with human evaluations(Chiang and Lee, 2023; Dubois et al., 2023; Liu et al.). Following this paradigm, we adopt GPT-4o and DeepSeek-R1 to score reconstructed articles relative to human-written originals. We employ a custom 1–5 scale rubric covering six key aspects: Relevance, Structure, Coverage, Accuracy, Coherence, and Overall Quality. Table 5 presents the detailed grading rubric.

B Pseudo Code

In §3, we present the complete pipeline of our framework, which consists of two major stages: Question-Focus Discourse Structuring and Argument Reconstruction, the latter comprising both Evidence Collection and Segment-Level Generation. Algorithm 1 displays the overall workflow of our work.

We implement the entire pipeline in a zero-shot prompting manner using the DSPy framework (Khattab et al., 2023). Detailed prompt configurations are shown in Listings 1, 2 and 3.

Algorithm 1 Question-Focus Discourse structuring and Argument Reconstruction

Input: Human-written article D
Output: Reconstructed article $\hat{D}$

```
// Discourse Structuring
T, P, B ← extract_metadata(D)
[AU_1, AU_2, ..., AU_n] ← segment_argument_units(D)
for each AU_i in [AU_1, ..., AU_n] do
    Q_i ← gen_guiding_question(AU_i, T, B)
    f_i ← extract_attentional_focus(AU_i)
end for
F ← {(AU_i, Q_i, f_i)}_{i=1}^n
// Evidence Collection
for each (AU_i, Q_i, f_i) in F do
    C_i^fact ← extract_factual_claims(AU_i)
    C_i^struct ← gen_queries_from_structure(Q_i, f_i, T, P, B)
    C_i ← C_i^fact ∪ C_i^struct
    R_i ← retrieve_articles(C_i)
end for
// Argument Reconstruction
for each (AU_i, Q_i, f_i, R_i) do
    [q_i1, q_i2, ...] ← decompose_subquestions(Q_i, f_i, T, B)
    snippets ← retrieve_snippets(q_ij, R_i)
    ÂU_i ← generate_segment(Q_i, f_i, snippets)
end for
Î, Ĉ ← generate_intro_conclusion(T, P, B)
D̂ ← assemble_article(Î, {ÂU_i}_{i=1}^n, Ĉ)
return D̂
```

```python

class ExtractMetaPrompt(dspy.Signature):
    """
    You are an expert in argument analysis.
    Given an article, your task is to extract the following three elements:
    1. Research Topic: the main issue or subject the article focuses on.
    2. Core Problem: the central problem the article aims to address or argue.
    3. Background Information: relevant contextual or factual details that help explain the topic and the core problem.
    Follow this format exactly:
    1. Research Topic:
    2. Core Problem:
    3. Background Information:
    """
    article = dspy.InputField(prefix="Article Content:\n", format=str)
    topic = dspy.OutputField(prefix="Research Topic:\n")
    core_problem = dspy.OutputField(prefix="Core Problem:\n")
    background = dspy.OutputField(prefix="Background Information:\n")

class ExtractGuidingQuestion(dspy.Signature):
    """
    You are an expert in argument structure analysis.
    Given the research topic, core problem, background information of the article, and a specific argument unit, please clearly identify the purpose of this argumentative unit, i.e., what it aims to argue and what question it seeks to answer.

    Format your response as follows:
    Guiding Question:
    """
    topic = dspy.InputField(prefix="Research Topic of the Article:\n", format=str)
    background = dspy.InputField(prefix="Background Information of the Article:\n", format=str)
    core_problem = dspy.InputField(prefix="Core Problem of the Article:\n", format=str)
    argument_unit = dspy.InputField(prefix="Content of a Specific Argumentative Unit in the Article:\n", format=str)
    guiding_question = dspy.OutputField(prefix="Guiding Question:\n")

class ExtractAttentionalFocus(dspy.Signature):
    """
    You are an expert in argument analysis.
    Given the research topic, core problem, background information, and the content of a specific argument unit, your task is to identify the main analytical perspectives or angles that this unit focuses on during the reasoning process.

    Format your response as follows:
    Focus 1:
    Focus 2:
    ...
    Focus n:
    """

    topic = dspy.InputField(prefix="Research Topic of the Article:\n", format=str)
    background = dspy.InputField(prefix="Background Information of the Article:\n", format=str)
    core_problem = dspy.InputField(prefix="Core Problem of the Article:\n", format=str)
    argument_unit = dspy.InputField(prefix="Content of a Specific Argumentative Unit in the Article:\n", format=str)
    attentional_focus = dspy.OutputField(prefix="Attentional Focus:\n")
```

Listing 1: Prompts used in our framework, corresponding to Line 2, 5, 6 in Algorithm 1.

```python
class ExtractEvidenceItems(dspy.Signature):
    """
    You are an expert in argument extraction.
    Based on the research topic, core problem, background information, and the content of a specific argument unit, your task is to identify all evidence used to support the argument in that unit.
    Evidence may include, but is not limited to:
    - Facts: objective statements or commonly accepted knowledge
    - Data: statistics, survey results, research findings, etc.
    - Events: real-world historical, social, or contemporary cases
    - Examples: specific and representative instances or scenarios
    - Other relevant types of support

    Extract all relevant evidence comprehensively. Each item should be a complete sentence taken directly from the original text. Present one piece of evidence per line, preserving the original wording.

    Format your response as follows:
    Evidence 1:
    Evidence 2:
    ...
    """

    topic = dspy.InputField(prefix="Research Topic of the Article:\n", format=str)
    background = dspy.InputField(prefix="Background Information of the Article:\n", format=str)
    core_problem = dspy.InputField(prefix="Core Problem of the Article:\n", format=str)
    argument_unit = dspy.InputField(prefix="Content of a Specific Argumentative Unit in the Article:\n", format=str)
    evidences = dspy.OutputField(prefix="Extracted Evidence:\n")

class GenerateSearchQueriesPrompt(dspy.Signature):
    """
    Your task is to generate a set of high-quality search queries based on the provided information.  These queries will be used with a search engine (e.g., Google) to find relevant materials or evidence supporting a specific argumentative issue.
    Each query should be focused on the guiding question and reflect its attentional focus.

    Please ensure the queries meet the following criteria:
    1. Be specific and targeted: avoid overly broad or generic keyword combinations.
    2. Prefer question formats: such as "How...", "Why...", or "What is the impact of...".
    3. Incorporate all input information: including the research topic, guiding question, background information, and key attentional focus areas.

    Format your response as follows:
    1. Query 1
    2. Query 2
    ...
    n. Query n
    """

    topic = dspy.InputField(prefix='Research Topic of the Article:\n', format=str)
    background = dspy.InputField(prefix='Background Information of the Article:\n', format=str)
    question = dspy.InputField(prefix='The Question the Argumentative Unit Aims to Answer:\n', format=str)
    attentional_focus = dspy.InputField(
        prefix='Attention focus for the Argumentative Unit (provided in list form):\n', format=list)
    queries = dspy.OutputField(prefix='Generated Search Queries:\n')
```

Listing 2: Prompts used in our framework (continue), corresponding to Line 11, 12 Algorithm 1.

```python
class GenerateSubQuestionsPrompt(dspy.Signature):
    """
    You are a professional research assistant.
    Based on a guiding question, the research topic, background information, and a set of attentional focus points, your task is to generate multiple additional sub-questions.

    Requirements:
    1. These sub-questions should help guide the collection of high-quality information to support argumentative analysis.
    2. Each sub-question should be closely aligned with the given focus areas, as they represent key angles for addressing the guiding question.
    3. The sub-questions should contribute to a deeper understanding and more precise elaboration of the guiding question.

    Format your output as follows:
        Question 1:
        Question 2:
        ...
        Question n:
    """

    topic = dspy.InputField(prefix="Research Topic:\n", format=str)
    question = dspy.InputField(prefix="Main Guiding Question:\n", format=str)
    attentional_focus = dspy.InputField(prefix="Attentional Focus (as a list):\n", format=list)
    background = dspy.InputField(prefix="Background Information:\n", format=str)
    sub_questions = dspy.OutputField(prefix="Generated Sub-Questions:\n", format=str)

class GenArgumentUnitPrompt(dspy.Signature):
    """
    You are an expert in argumentative writing.
    Based on the research topic, core problem, background information, guiding question, attentional focus, and collected evidence, write a well-reasoned, and evidence-based argument unit.

    Requirements:
    1. Focus on the guiding question and attentional focus. Interpret the input with clear purpose and reasoning.
    2. Analyze each focus area in depth. Avoid surface-level descriptions.
    3. Ensure accuracy, avoid redundancy, and do not fabricate content.

    """
    topic = dspy.InputField(prefix="Research Topic of the Article:", format=str)
    background = dspy.InputField(prefix="Background Information of the Article:", format=str)
    core_problem = dspy.InputField(prefix="Core Problem of the Article:", format=str)
    question = dspy.InputField(prefix="Guiding Question of the Argumentative Unit:", format=str)
    attentional_focus = dspy.InputField(
        prefix='Attentional focus for the Argumentative Unit (provided in list form):', format=list)
    context = dspy.InputField(prefix="Collected Relevant Evidence:\n", format=str)
    output = dspy.OutputField(prefix="Generated Argumentative Unit Content:\n", format=str)
```

Listing 3: Prompts used in our framework (continue), corresponding to Line 18, 20 Algorithm 1.

Criteria Description	**Relevant:** Assesses how well the reconstructed article aligns with the original in themes, claims, and key information.
Score 1 Description	Major inconsistencies, misrepresenting the original core ideas.
Score 2 Description	Some deviations or missing information, but the main ideas are still conveyed.
Score 3 Description	Generally consistent, with some deviations in details but core ideas intact.
Score 4 Description	Mostly consistent, minor differences that don't affect the core content.
Score 5 Description	Fully aligned with the original, with only minor differences that don't affect understanding.
Criteria Description	**Structure:**Assesses how accurately the article preserves the original structure and logic.
Score 1 Description	Severe structural misalignment, lacking logical flow.
Score 2 Description	Significant structural deviations, major themes present but sub-dimensions misaligned.
Score 3 Description	Structure generally aligned, but some sub-dimensions deviated or omitted.
Score 4 Description	Mostly preserves the structure, with minor adjustments that don't affect the flow
Score 5 Description	Fully preserves the original structure and logic, with accurate themes and sub-dimensions.
Criteria Description	**Coverage:** Assesses the extent to which the article covers key points and information from the original.
Score 1 Description	Major points and key information missing, incomplete content.
Score 2 Description	Some key points missing, but core ideas still conveyed.
Score 3 Description	Covers most key points, but some details or secondary information are missing.
Score 4 Description	Covers most key points, with minor omissions that don't affect understanding.
Score 5 Description	Comprehensive coverage of all major points and key information.
Criteria Description	**Accuracy:** Assesses the accuracy of key facts, arguments, and data referenced in the reconstructed article.
Score 1 Description	Major errors that undermine the article's accuracy.
Score 2 Description	Several inaccuracies that affect the article's credibility.
Score 3 Description	Some inaccuracies, but overall impact is minimal.
Score 4 Description	Most facts are accurate, with minor errors that don't affect the overall content.
Score 5 Description	All facts, arguments, and data are fully accurate.
Criteria Description	**Consistency**:Assesses how accurately the article conveys the original's ideas, claims, and logic
Score 1 Description	Major inconsistencies, misrepresenting the original core ideas.
Score 2 Description	Some deviations or missing information, but the main ideas are still conveyed.
Score 3 Description	Generally consistent, with some deviations in details but core ideas intact.
Score 4 Description	Mostly consistent, minor differences that don't affect the core content.
Score 5 Description	Fully aligned with the original, with only minor differences that don't affect understanding.
Criteria Description	**overall:** Assess the overall quality of the reconstructed article by assigning a score from 1 to 5, reflecting its fidelity to the original content across all relevant dimensions, including but not limited to content relevant , structural integrity, information coverage, Content Accuracy, and Semantic Consistency.

Table 5: Scoring rubrics on a 1-5 scale for the evaluator LLM.

Argument Unit 1 (human-write)：

其次是国内因素驱动。特朗普的支持者基本盘中，基督教福音派是核心力量之一。福音派对以色列的狂热支持具有深厚宗教基础。福音派信徒从宗教视角审视以色列“在迦南地建国”的意义，将对以色列的支持看作是践行宗教义务的一部分。许多福音派信徒公然反对巴勒斯坦建国计划，反对“两国方案”，支持以色列的定居点扩建计划，并鼓吹以色列应占领全部巴勒斯坦地区。事实上，特朗普任命的美国中东问题特使维特科夫以及候任驻以色列大使赫卡比在相关问题上的立场就强烈地体现出福音派世界观和对以色列的绝对支持。特朗普目前极具争议的加沙治理方案显示出美国国内右翼力量在巴勒斯坦问题上的巨大话语影响力。

Argument Unit 1 (reconstruct)：

特朗普关于加沙治理的方案在国内之所以能够获得一定支持，主要源于美国国内政治因素的复杂作用。特朗普的方案受到了部分美国右翼势力的强力支持，尤其是基督教福音派。他们不仅是特朗普的重要支持者，也因宗教信仰和政治立场的原因普遍反对巴勒斯坦的建国计划，认为以色列在地缘政治上是美国不可或缺的盟友。这些群体对中东事务持强烈的偏袒以色列立场，而特朗普的政策迎合了这一群体的核心诉求。此外，特朗普政府内的关键官员，如中东问题特使维特科夫和候任驻以色列大使赫卡比，均对以色列表现出强烈支持。这些官员的立场反映了特朗普政府对以色列政策的态度，也为其加沙治理方案提供了内部政策上的支持和执行上的保障。在政治运作上，特朗普借助这些人力资源和外交政策的影响力，不仅稳固了其在共和党内部的领导地位，也进一步争取了以色列在美国国内的支持者，形成了一种正反馈机制，从而加强了对其方案的国内政治支持。

Argument Unit 2 (human write)：

特朗普政府推出的“战后加沙治理方案，首先是地缘政治考量。无论是将以色列总理内塔尼亚胡作为特朗普再度就任美国总统后首位接待的外国领导人，还是为全部出口以色列武器放行，目前来看，延续第一任期的对以全面偏袒政策仍将成为特朗普本届政府中东政策的基础底色。近十余年来，美国不断在中东实施战略收缩，但又想维持基于硬实力的军事威慑，力图以低成本维系在中东的影响力，因此就必须塑造一个可靠的战略支点，以色列无疑是绝佳选择。在巴勒斯坦问题上，以色列政府立场愈发强硬，拒绝了巴勒斯坦方面、阿拉伯国家以及国际社会提出的多种战后加沙治理方案。特朗普此时提出“接管加沙”，不仅为以色列解套，更拓展了美以同其他各方在地缘政治角力中讨价还价的“绝佳”空间。

Argument Unit 2 (reconstruct)：

特朗普提出加沙治理方案的背后，充满了复杂的地缘政治考量。首先，此举无疑是对美国在中东地区影响力的再次确立和延续。尽管美国近年来在中东采取战略收缩政策，试图以低成本维持其硬实力的军事威慑，但特朗普政府仍力图塑造以色列为可靠的战略支点。特朗普在任期间，首次接待的外国领导人即是以色列总理内塔尼亚胡，并为出口以色列的武器全面放行，显示出对以色列的高度偏袒，这种政策延续至他的加沙方案中。此外，该方案反映了其试图通过极端的地缘政治手段获取更多谈判筹码的意图。美国对加沙的接管计划，虽然在国际法和实际操作层面几乎难以实现，但此举却可能为美国在中东和平进程中增加谈判优势。特朗普的战略似乎在于，通过提出极具争议的方案，迫使国际社会在一个相对有利于美国的立场上妥协，从而实现其在中东地区更大的话语权。

Figure 3: Example of a human-written argumentative unit and a structure-guided argument reconstruction unit generated based on our proposed framework.